Table 1: Relevant files and descriptions in Bioinformatics\_Group\_Project

|  |  |
| --- | --- |
| **Description** | **File name** |
| BLAST results | alignmenthittable[1,2,6,7,8,9,10].csv |
| Top BLAST hits | alignfinal.txt |
| Script for getting top blast hit | align.sh |
| Fasta files for 10 blast results for top hits in alignment | sequence[1,2,6,7,8,9,10].fasta |
| Script for muscle build, search, and hits | HMM\_script.sh |
| Muscle alignments | sequence[1,2,6,7,8,9,10].txt |
| HMM builds | sequence[1,2,6,7,8,9,10].hmm |
| HMM searches | sequence[1,2,6,7,8,9,10]\_[Control[1-2]protein.fasta\_cl, Obese[1-2]protein.fasta\_cl].txt |
| Short answer primate sequences | sequence[1,6]\_primate.fasta |
| Short answer nonmammal sequences | sequence[1,6]\_nonmammal.fasta |
| Muscle alignments | sequence[1,6]\_primate.txt |
| Muscle alignments | sequence[1,6]\_nonmammal.txt |
| HMM builds | sequence[1,6]\_primate.hmm |
| HMM builds | sequence[1,6]\_nonmammal.hmm |
| HMM searches | sequence[1,6]\_primate\_[Control[1-2]protein.fasta\_cl, Obese[1-2]protein.fasta\_cl].txt |
| Total hits in searches | HMM\_hits.txt |
| Individual hit tables for graphing | sequence[1,2,6,7,8,9,10]\_hits.txt |
| Script for bar graphs of hits | biogroupproject.py |
| Final bar graphs for each sequence | sequence[1,2,6,7,8,9,10]hits.pdf |

**Qualitatively compare expression level results to those reported in Kunz and Pluznick 2017**

This paper graphically shows the difference in gene expression in healthy versus obese mice in 11 genes. We looked at 6 genes. The Kunz and Pluznick figure 3 graphs show the ratio of gene expression of the high fat diet mice to the control mice through RNA seq results. For our sequence 1, glutathione S transferase, the expression in obese mice was downregulated compared to controls. This matched the paper’s results for Gsta2. Our sequence 2,

Slc7a12, had ambiguous results, but the obese mice downregulated Slc7a12 compared to control 2. The paper also showed a downregulation in the obese mice compared to controls by 10%. Sequence 6, Ptpn5, also had ambiguous results. Obese expression compared to control 1 showed a two-fold increase, while there was no difference compared to control 2. The paper, on the other hand, found a 10-fold increase in expression in the obese mice. Sequence 8, Atp12a, showed a two-fold increase in expression in the Obese mice compared to controls. In the paper, they showed a 22-fold increase in expression. Sequence 9, Lhx2, showed a three-fold increase in expression in the Obese mice compared to controls. In the paper, they showed a 28-fold increase in expression. Sequence 10, Synpr, showed an 8-fold increase in expression in the Obese mice compared to controls. In the paper, they showed Synpr upregulated 43-fold in the obese mice.

Short answers

**1.**  For transcript 1, when limited to mus musculus, the blast search results from megablast have 100 hits with identity scores ranging from 100% to 76%, e values from 0 to 4e-20, and max scores and total scores from 1528 to 106. The discontinuous megablast results have 100 hits with identity scores ranging from 100% to 69%, e values from 0 to 3e-4, and max and total scores from 1541 to 53.6. The blastm results have 100 hits with identity scores ranging from 100% to 69%, e values from 0 to 7.2, and max scores and total scores from 1541 to 39.2. For transcript 2, when limited to mus musculus, the blast search results from megablast have 15 hits with identity scores ranging from 100% to 82%, e values from 0 to 3e-145, and max scores and total scores ranging from 3236 to 521. The discontinuous megablast results have 24 hits with identity scores ranging from 100% to 64%, e values from 0 to 4e-9 and max and total scores from 3160 to 70.7. The blastm results have 100 hits with identity scores ranging from 100% to 64%, e values from 0 to 2.1, and max and total score from 3160 to 41. The megablast search returns the most highly similiar sequences so these results will always have the highest identity scores, lowest e values, and highest max and total scores. In contrast, the discontinuous megablast will retrieve more dissimiliar sequences than the megablast, so there will likely be more hits, if the megablast search did not have 100 hits to begin with, and the range of e values, identity scores, and total score and max score will be larger. Blastm returns the most dissimiliar sequences of all the algorithms resulting in the most hits and the largest range of e values, identity scores, and total and max scores.

**2.** We took human and drosophila sequences from blast for sequence 1 protein. Compared to the mouse hmm builds, all mouse, human, and drosophila had the same number of “hits” or expression level in the controls and obese mice. However, their e-values expectedly differed slightly. E-values represent the random background noise or number of hits expected by chance, therefore lower e-values mean better identity scores between alignments. The mice proteins had the lowest e-values, followed by human proteins, then drosophila. We suspect this is a highly conserved protein across species. We also took human and alligator sequences from blast for sequence 6 protein. Compared to the mouse hmm builds, the human and alligator sequences generated no hits or no expression in the controls and obese mice. Perhaps this is a nonconserved protein which may have species-specific sequences.

Sequence 1 control 1: e value: 2.2e-112, hits=12

Sequence1 control 2: e value: 3e-112, hits=20

Sequence 1 obese1: e value: 6.8e-12, hits=2

Sequence1 obese2: e value: 6.5e-112, hits=1

Sequence1 primate control 1: e value: 1.1e-83, hits=12

Sequence 1 primate control 2: e value: 1.5e-83, hits=20

Sequence 1 primate obese 1: e value: 3.3e-83, hits=2

Sequence 1 primate obese 2: e value: 3.2e-83, hits=1

Sequence 6 control 1: e value: 0, hits=12

Sequence 6 control 2: e value: 0, hits=20

Sequence 6 obese 1: e value: 0, hits=20

Sequence 6 obese 2: e value: 0, hits=18

Sequence 1 nonmammal control 1: e value 4.7e-17, hits=12

Sequence 1 nonmammal control 2: e value 6.4e-17, hits=20

Sequence 1 nonmammal obese 1: e value 1.5e-16, hits =2

Sequence 1 nonmammal obese 2: e value 1.4e-16, hits=1